

**ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ**

**Федеральное государственное автономное  
образовательное учреждение высшего образования**

**Национальный исследовательский университет  
«Высшая школа экономики»**

Факультет экономических наук  
Образовательная программа «Экономика»

**КУРСОВАЯ РАБОТА**

На тему «Вариация алгоритма кросс-валидации со взвешиванием наблюдений»

Студент группы БЭК161  
Гармидер Петр Александрович

Научный руководитель:  
Борис Демешев

Москва 2019

# Содержание

<b>1</b>	<b>Введение</b>	<b>3</b>
1.1	Параметры и гиперпараметры модели . . . . .	3
1.2	Типы кросс-валидации . . . . .	3
1.3	Кросс-валидация для временных рядов . . . . .	6
<b>2</b>	<b>Взвешенная кросс-валидация для временных рядов</b>	<b>8</b>
2.1	Мотивация . . . . .	8
2.2	Кросс-валидация для временных рядов со взвешиванием наблюдений . . . . .	9

# 1 Введение

## 1.1 Параметры и гиперпараметры модели

Большая часть популярных моделей имеют множество параметров и гиперпараметров однозначно выделяющие модель из множества всех алгоритмов. Гиперпараметры модели — параметры, вводимые пользователем вручную, которые в большинстве случаев не меняются в ходе обучения<sup>1</sup>. Параметры модели — параметры, которые не вводятся пользователем вручную, а есть результат оптимизации функции потерь. Их конечные значения становятся известны после завершения процесса обучения. Параметры модели связывают имеющиеся у исследователя данные и выбранный алгоритм для обучения.

Рассмотрим пример. Пусть имеем обучающую выборку  $D = \{x_i, y_i\}$  и тестовую выборку  $d = \{x_i, y_i\}$ , где  $x_i, y_i \in R$ ,  $|D| = N$  и  $|d| = n$ . Будем оценивать  $y_i$  используя  $L_2$  - регуляризатор или Ridge-regression (см. [2]). Имеем безусловную задачу оптимизации:

$$\sum_{i=1}^N (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 + \lambda \hat{\beta}_1^2 \rightarrow \min_{\hat{\beta}_0, \hat{\beta}_1} \quad (1)$$

В (1) существуют два параметра для оптимизации  $\hat{\beta}_0$  и  $\hat{\beta}_1$ , а также один гиперпараметр  $\lambda$ , который является константой в рассматриваемой задаче. Коэффициенты регрессии находятся из задачи минимизации при заданном уровне  $\lambda$ , в то время, как степень регуляризации  $\lambda$  задается исследователем. В конечном счете исследователь заинтересован в создании модели, которая имеет высокое качество на тестовой выборке  $d$ , что не участвовала в процессе обучения. Качество работы модели на выборке  $d$  очевидно зависит от выбранного исследователем  $\lambda$ . Поэтому, выбор оптимального значения для гиперпараметров является также важной задачей для получения лучшей модели.

Гиперпараметрами модели также могут быть: метод обработки пропусков в данных, количество слоев в нейронной сети, выбранные функции активации, уровень Dropout, скорость обучения и другие.

Влияние гиперпараметров на качество работы модели делает их правильный выбор отдельной задачей. Оптимальный алгоритм подбора гиперпараметров уникален для каждой конкретной задачи. Качество работы модели в целом принято оценивать на выборках не участвовавших в обучении, но для которых известно истинное значение зависимой переменной.

## 1.2 Типы кросс-валидации

Кросс-валидация — техника валидации модели для оценки качества её работы и установления факта обобщающей способности. Метод CV позволяет понять, выучила ли модель зависимость между рассматриваемыми переменными или же алгоритм переобучился и хорошо предсказывает

---

<sup>1</sup>Однако такая практика также используется, например, изменение гиперпараметра learning rate в ходе алгоритма градиентного спуска (см. [1])

лишь данные имеющиеся в обучающей выборке<sup>2</sup>. Как правило, кросс-валидационной проверке подвергаются модели, которые направлены на точность предсказаний, оставляя за бортом вопрос интерпретации полученных результатов.

Разберем некоторые варианты алгоритма кросс-валидации:

Пусть имеем:  $X$  — множество признаков, описывающих объекты;  $Y$  — множество зависимых переменных;  $D^l = \{x_i, y_i\}$  — наблюдаемая выборка, где  $x_i \in X$ ,  $y_i \in Y$ ,  $l$  — размер выборки;  $Q : (A \times (X \times Y)) \rightarrow R$  — функция потерь;  $A$  — модель;  $\mu : (X \times Y) \rightarrow A$  — алгоритм обучения [3].

- Валидация на отложенных данных (Hold-out)

Исследователь выбирает число  $t$  — количество объектов из множества  $D$ , которые будут использованы для обучения модели. Соответственно, оставшая часть объектов  $l - t$  используется для проверки качества работы модели. Итого, получаются две выборки  $Train^t$  и  $Test^{l-t}$  такие, что  $Tr^t \cup T^{l-t} = D^l$ . Решается задача:

$$HOCV(\mu, Tr^t, T^{l-t}) = Q(\mu(Tr^t), T^{l-t}) \rightarrow \min_{\mu}$$

$Tr^t$	$T^{l-t}$
--------	-----------

Рис. 1: Иллюстрация валидации на отложенных данных

Метод Hold-out CV обычно используется в том случае, если исследователь обладает большой обучающей выборкой, т.к. данный способ не требует больших вычислительных мощностей. Итоговое качество  $Q$  также зависит от разбиения обучающей выборки, что является главным недостатком метода.

- K-fold кросс-валидация

Исходная выборка  $D$  разбивается случайным образом на  $K$  непересекающихся, примерно равных по мощности множеств:  $D_1, D_2, \dots, D_k$ ;  $|D_i| \approx \frac{l}{k}$ . После чего, для каждого из получившихся множеств проводится процедура hold-out CV; результаты всех процедур усредняются. Решается задача:

$$KFCV(\mu, D, K) = \frac{1}{K} \sum_{i=1}^K OFCV(\mu(D \setminus D_i), D_i) \rightarrow \min_{\mu}$$

Метод K-fold CV решает проблему высокой зависимости получаемого результата от разбиения, однако является весьма затратным с точки зрения вычислительных мощностей.

Обычно используется в тех случаях, когда размеры выборки и модель позволяют быстро

<sup>2</sup>В английской литературе данную ситуацию называют overfitting

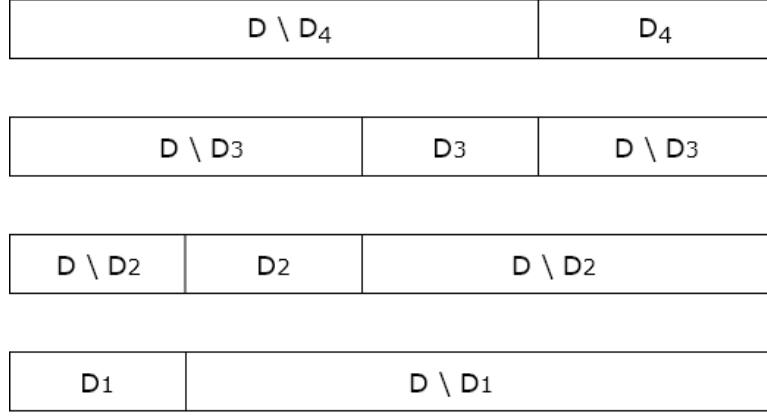


Рис. 2: Иллюстрация K-fold кросс-валидации при  $K = 4$

проводить процедуру обучения. Число  $K$  выбирается исследователем на своё усмотрение. Заметим, что при  $t = \frac{l}{2}$  Hold-out CV  $\equiv$  two-fold CV.

- Leave-one-out кросс-валидация

Частный случай K-fold кросс-валидации, при  $K = l$ . Исходное множество  $D$  разбивается на  $l$  подмножеств:  $D_1, D_2, \dots, D_l$ ;  $|D_i| = 1$ . После чего проводится стандартная K-fold кросс-валидация. LOO CV подвергается критике; в некоторых исследованиях говорится, что данный метод плохо оценивает предсказательную силу модели [4]. Кроме того, данный метод требует высоких вычислительных мощностей, т.к. потребуется  $l$  раз обучать модель.

- Полная кросс-валидация (Complete)

Исследователь выбирает число  $t$ , после чего изначальная выборка  $D^l$  разбивается всеми возможными способами на выборки  $Tr^l$  и  $Tt^{l-t}$ . Заметим, что число возможных разбиений для заданного  $t$  равно  $C_l^{l-t}$ . Таким образом, исследователь решает задачу:

$$CCV(D, t) = \frac{1}{C_l^{l-t}} \sum_{D^l = Tr^l \cup Tt^{l-t}} Q(\mu(Tr^t), Tt^{l-t}) \rightarrow \min_{\mu}$$

Даже при достаточно небольших значениях  $t$ , данный метод проверки работоспособности модели используется крайне редко в силу его вычислительной сложности.

- Случайные разбиения (Random sampling)

Выборка разбивается в случайной пропорции, после чего для получившегося разбиения проводится hold-out CV. Данная процедура повторяется несколько раз; результаты усредняются.

- $M \times K$ -fold кросс-валидация

K-fold кросс-валидация проводится  $M$  раз; результаты  $M$  валидаций усредняются. Итого:

$$MKFCV(\mu, D, K, M) = \frac{1}{M} \sum_{i=1}^M KFCV(\mu, D, K) \rightarrow \min_{\mu}$$

Метод допустим к использованию при небольшой выборке и алгоритме, который способен обучаться  $M \times K$  раз за разумное время.

Одной из причин использования кросс-валидации является подбор гиперпараметров. Допустим исследователь решает задачу по выбору оптимального параметра  $\lambda$  для  $L_2$  - регуляризатора на странице 3. Если не проводить кросс-валидацию, то лучшее качество на выборке  $D$  даст модель с  $\lambda = 0$ , т.к в этом случае на коэффициенты регрессии не накладываются никакие ограничения, а поэтому они сойдутся к решению, которое минимизирует среднеквадратичную ошибку на выборке  $D$ . Однако, получаемое качество работы модели на обучающей выборке не отображает реальную картину. Исследователь, в конечном счете, заинтересован создать модель, которая будет иметь высокое качество на объектах не входящих в обучающую выборку. Для этого, есть смысл проверять качество работы модели используя один из перечисленных методов кросс-валидации. В этом случае, оптимальное значение  $\lambda$  скорее всего окажется больше нуля.

### 1.3 Кросс-валидация для временных рядов

Временной ряд — наблюдаемая выборка данных, объекты которой представлены во временном порядке. В большинстве случаев, временной ряд представлен точками, которые одинаково отдалены друг от друга во временной шкале. Анализ временных рядов значимо отличается от подходов работы с простой выборкой данных: учитывается зависимость точек от времени, а также взаимосвязь текущего значения параметра с лагированными. Примерами временных рядов являются: курс доллар к евро, реальный уровень ВВП США, ключевая ставка ЦБ РФ, последовательность кадров в видеоролике<sup>3</sup> и так далее. В данной работе фокус будет сделан на одномерных временных рядах, прогноз для которых будет иметь вид некоторой функции от прошлых значений.

Как и раньше, задача построения моделей для прогнозирования временных рядов включает в себя стадию подбора оптимальных гиперпараметров модели. Однако, по очевидным причинам большинство методов кросс-валидации из секции 1.2 не подходят для случая, когда объектом исследования является временной ряд. Непрактично оценивать модель на случайно выбранных данных, которые с большой долей вероятности потеряют временную структуру при разбиениях.

Очевидным решением в таком случае является разбиение исходной выборки на две части: обучающую и валидационную — с сохранением временной структуры. После чего на последней измерять качество работы модели. Тем не менее, проблема высокой зависимости результата от разбиения становится вновь актуальной. Возможно, исследователь столкнется с ситуацией, когда

---

<sup>3</sup>Объектами в таком случае являются трехмерные матрицы, хранящие в себе значения, характеризующее интенсивность RGB каналов

качественная модель плохо справляется лишь с отведенным для валидации блоком данных, но будет давать прогнозы с высокой точностью в долгосрочной перспективе.

Решением данной проблемы является адаптированный алгоритм K-fold кросс-валидации для временных рядов (K-fold TSCV):

Пусть  $TS = \{y_t\}$  — наблюдаемая выборка временного ряда, где  $|TS| = T$ ,  $t = 1, 2, \dots, T$ ;  $\lambda \in \Lambda$ , где  $\Lambda$  — множество всех доступных гиперпараметров для модели  $\mu$ . Исследователь выбирает число  $K$ , после чего  $TS$  делится на  $K$  блоков:  $TS_1, TS_2, \dots, TS_K$ ;  $|TS_i| \approx \frac{T}{K}$ ;  $TS_i = \{y_j, y_{j+1}, y_{j+2} \dots\}$ . Затем, выбранная модель  $\mu$  обучается  $K - 1$  раз следующим образом: обучение проходит на  $TS_1$  — качество проверяется на  $TS_2$ , обучение проходит на  $TS_1 \cup TS_2$  — качество проверяется на  $TS_3$  и так далее. Для подбора оптимального гиперпараметра  $\lambda$  для модели  $\mu$  исследователь решает задачу:

$$TSCV_K = \frac{1}{K-1} \sum_{i=1}^{K-1} Q(\mu(TS_1 \cup \dots \cup TS_i, \lambda), TS_{i+1}) \rightarrow \min_{\lambda \in \Lambda} \quad (2)$$

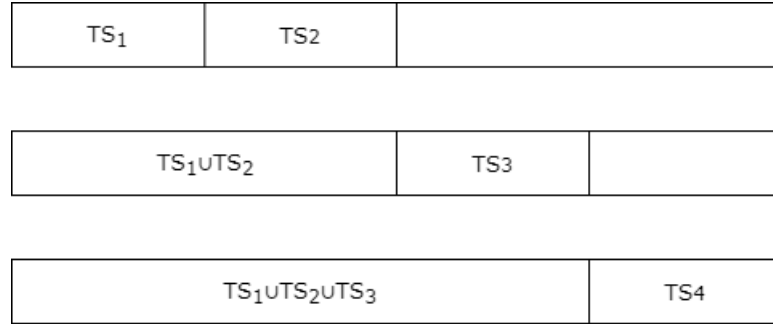


Рис. 3: Иллюстрация K-fold кросс-валидации для временных рядов при  $K = 4$

Описанный подход также не является идеальным решением и плохо работает в случаях, когда наблюдений не так много, а следовательно, алгоритму на первых итерациях предстоит обучаться на малом количестве наблюдений и измерять качество на такой же по количеству выборке. Очевидно, что в таких ситуациях модель покажет низкое качество, но это не значит, что она не применима для качественного прогнозирования рассматриваемого ряда.

## 2 Взвешенная кросс-валидация для временных рядов

### 2.1 Мотивация

Зависимость временного ряда во времени делает его весьма интересным объектом для изучения. Как правило, рассматриваемые в прикладных задачах ряды являются объектами человеческой деятельности: продажи товаров, количество новых клиентов, — что по определению делает такие ряды подверженными внешним шокам. Хотелось бы подобрать такой алгоритм, который устойчив к данным изменениям и способен хорошо предсказывать будущие значения по всей имеющейся информации на сегодняшний день. В некоторой степени данную задачу может разрешить алгоритм ETS с аккуратно подобранными гиперпараметрами. Ввиду особенностей своей работы ETS позволяет моделировать, например, изменение долгосрочного уровня в ответ на произошедшие в прошлом шоки. Однако, дабы такая модель была решением задачи поиска оптимального алгоритма на кросс-валидации для временных рядов, ETS также должна хорошо предсказывать прошлые данные в то время, когда исследователя, как правило, интересуют актуальные наблюдения. Данные рассуждения наталкивают на идею модификации алгоритма кросс-валидации для устранения упомянутой проблемы.

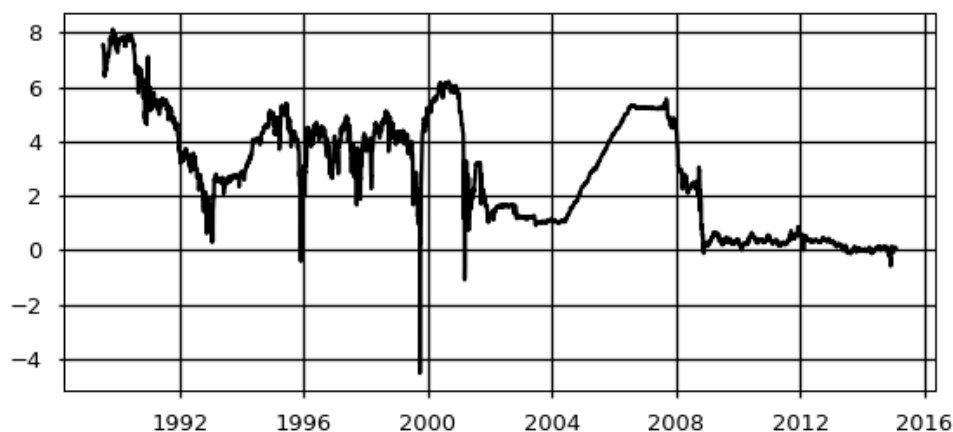


Рис. 4: Месячный форвардный курс золота (GOFO), в процентах [5]

Чтобы показать, что ситуация описанная выше не является теоретизированной, рассмотрим рис. 4. На данном графике изображен месячный форвардный курс золота в процентах от текущей стоимости за период с 1990 по начало 2015 года. Довольно очевидно, учитывая имеющиеся данные за последние 8 лет, что сильным и, возможно, наилучшим прогнозом для рассматриваемого ряда на ближайшее время будет векторов скаляров в окрестности нуля. Однако, маловероятно, что такая модель будет отобрана на кросс-валидации, учитывая поведение ряда до 2009 года. Видно, что ряд имел участки, где присутствуют: линейно-возрастающий тренд; линейно-убывающим тренд; ярковыраженная сезонность и другие особенности, которые не свойственны



актуальным наблюдениям. Некоторый класс моделей справляются и с такими особенностями ряда. Однако, хотелось бы адаптировать процесс кросс-валидации так, чтобы практически любая модель имела устойчивость к изменчивости характеристик временного ряда, при процессе подбора гиперпараметров

## 2.2 Кросс-валидация для временных рядов со взвешиванием наблюдений

Основная идея за предлагаемым алгоритмом заключается в вере о том, что модель, которая качественно предсказывает последние наблюдения лучше той, что хорошо работает лишь на давних участках данных. Действительно, довольно часто исследователя вовсе не волнует ситуация, когда модель совершенно не справляется с давними участками, если алгоритм показывает высокое качество на актуальных. Данный факт обуславливается тем, что временные ряды часто используются для принятия решений в данный момент времени, основываясь на том, что будет завтра. Поэтому для пользователя важно иметь высокое качество в прогнозах именно на завтра<sup>4</sup>.

Отсюда вытекает обобщение существующего алгоритма кросс-валидации. Предлагается учитывать качество работы модели на недавних участках с большим весом, а качество работы модели в далеком прошлом с меньшим или не учитывать вовсе. Формально задача сводится к следующей:

Пусть  $TS = \{y_t\}$  — наблюдаемая выборка временного ряда, где  $|TS| = T$ ,  $t = 1, 2, \dots, T$ ;  $n(y)$  — номер наблюдения для  $y$ ;  $Q(a, b)$  — некоторая функция, которая измеряет различие  $a$  и  $b$ ;  $\mu(\lambda)$  — алгоритм прогнозирования;  $\lambda \in \Lambda$ , где  $\Lambda$  — множество всех доступных гиперпараметров для модели  $\mu$ . Проводим стандартную K-fold кросс-валидацию для временных рядов (2), сохраняя при этом качество работы модели для каждого объекта из отложенных выборок:  $TS_2, TS_3, \dots, TS_K$ . Итого, для алгоритма  $\mu(\lambda)$  имеем  $\approx \frac{(K-1)T}{K}$  предсказаний для объектов на отложенных выборках<sup>5</sup>. Данные предсказания составляют множество  $V(\mu(\lambda))$ . Тогда подбор оптимального гиперпараметра  $\lambda$  для алгоритма  $\mu$  сводится к следующей задаче минимизации:

$$\sum_{\hat{y} \in V(\mu(\lambda))} Q(\hat{y}, y) \gamma^{T-n(y)} \rightarrow \min_{\lambda \in \Lambda} \quad (3)$$

---

<sup>4</sup>Существуют задачи, целью которых является качественное прогнозирование прошлых периодов. Например, восстановление пропущенных данных о темпе роста ВВП нынешней РФ в период советской власти

<sup>5</sup> $TS_1$  требуется для минимального обучения модели

## Список литературы

- [1] Zeiler Matthew D. ADADELTA: An Adaptive Learning Rate Method. 2012.
- [2] Hoerl Arthur E, Kennard Robert W. Ridge regression: Biased estimation for nonorthogonal problems // Technometrics. 1970. Т. 12, № 1. С. 55–67.
- [3] Кросс-валидация. URL: <http://neerc.ifmo.ru/wiki/index.php?title=Кросс-валидация>.
- [4] Efron Bradley. How biased is the apparent error rate of a prediction rule? // Journal of the American statistical Association. 1986. Т. 81, № 394. С. 461–470.
- [5] Gold Forward Offered Rates (GOFO). URL: <https://www.quandl.com/data/LBMA/GOFO-Gold-Forward-Offered-Rates-GOFO>.